IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555
Vol. 2, No.3, June 2012

# Name Entity Recognition for Punjabi Language

Kamaldeep Kaur
University Institute of Engg. & Technology
Panjab University
Chandigarh, India

Vishal Gupta
University Institute of Engg. & Technology
Panjab University
Chandigarh, India

***Abstract*—This paper introduces Named Entity Recognition (NER) for Punjabi. Natural Language Processing (NLP) is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things. NER is a sub problem or application of NLP. Not much work has been done in NER for Indian languages in general and Punjabi in particular. Adequate annotated corpora are not yet available in Punjabi. The paper represents the Name Entity Recognition system for Punjabi language to seek and classify words which represent proper names in text into predefined categories like location, person-name, organization, date, time, designation etc. First we survey about the various approaches available for NER, then represent our hybrid approach for Punjabi. A number of language independent and language dependent features are extracted. The experimental results are shown.**

***Keywords-NLP, NER***

## I. INTRODUCTION

[1] Natural Language processing (NLP) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages. In theory, natural-language processing is a very attractive method of human-computer interaction. Natural-language understanding is sometimes referred to as an AI-complete problem, because natural-language recognition seems to require extensive knowledge about the outside world and the ability to manipulate it.

The history of NLP generally starts in the 1950s, although work can be found from earlier periods.

Although NLP may encompass both text and speech, work on speech processing has evolved into a separate field. Natural language generation systems convert information from computer databases into readable human language.

NLP is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things. NLP researchers aim to gather knowledge on how human beings understand and use language so that appropriate tools and techniques can be developed to make computer systems understand and manipulate natural languages to perform the desired tasks. The foundations of NLP lie in a number of disciplines, viz. computer and information sciences, linguistics, mathematics, electrical and electronic engineering, artificial intelligence and robotics, psychology, etc. [2]

[3]Research in natural language processing has been going on for several decades dating back to the late 1940s. Machine translation (MT) was the first computer-based application related to natural language.

Natural language processing approaches fall roughly into four categories: symbolic, statistical, connectionist, and hybrid. Symbolic and statistical approaches have coexisted since the early days of this field. Connectionist NLP work first appeared in the 1960's. For a long time, symbolic approaches dominated the field. In the 1980's, statistical approaches regained popularity as a result of the availability of critical computational resources and the need to deal with broad, real-world contexts. Connectionist approaches also recovered from earlier criticism by demonstrating the utility of neural networks in NLP.

Various sub problems in NLP include speech segmentation, text segmentation, part of speech tagging, word sense disambiguation, syntactic ambiguity.ent and are identified in italic type, within parentheses, following the example. Some components, such as multi-leveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

Major tasks in NLP include:

- Automatic summarization
- Foreign language reading aid
- Foreign language writing aid
- Information extraction
- Information retrieval (IR) - IR is concerned with storing, searching and retrieving information. It is a separate field within computer science (closer to databases), but IR relies on some NLP methods (for example, stemming). Some current research and applications seek to bridge the gap between IR and NLP.

- Machine translation - Automatically translating from one human language to another.
- Named entity recognition (NER) - Given a stream of text, determining which items in the text map to proper names, such as people or places. Although in English, named entities are marked with capitalized words, many other languages do not use capitalization to distinguish named entities.
- Natural language generation
- Natural language search
- Natural language understanding
- Optical character recognition
- Anaphora resolution
- Query expansion
- Question answering - Given a human language question, the task of producing a human-language answer. The question may be a closed-ended (such as "What is the capital of Canada?") or open-ended (such as "What is the meaning of life?").
- Speech recognition - Given a sound clip of a person or people speaking, the task of producing a text dictation of the speaker(s). (The opposite of text to speech.)
- Spoken dialogue system
- Stemming
- Text simplification
- Text-to-speech
- Text-proofing

## II. Name Entity recognition

### A. Introduction

Named entity recognition (NER) is a precursor for many natural languages processing tasks [4]. It is now firmly established as a key technology for understanding low- level semantics of texts [5]. It involves the identification of named entities such as person names, location names, names of organizations, monetary expressions, dates, numerical expressions etc. In the taxonomy of Computational Linguistics, NER falls within the category of Information Extraction which deals with the extraction of specific information from given documents [6].

[5] [7] [8] [9] [10] [18] The main role of NER is to identify expressions such as date and time as well as names of people, places, and organizations. Those expressions are difficult to extract using traditional natural language processing (NLP) because they belong to the open class of expressions, i.e. there is an infinite variety and new expressions are constantly being created. Automatically extracting proper names is useful to many problems such as machine translation, information retrieval, information extraction, question answering and summarization.

NER has important significance in the Internet search engines and in many of the Language Engineering applications. [6]

The architecture of a typical NER system can be shown as:

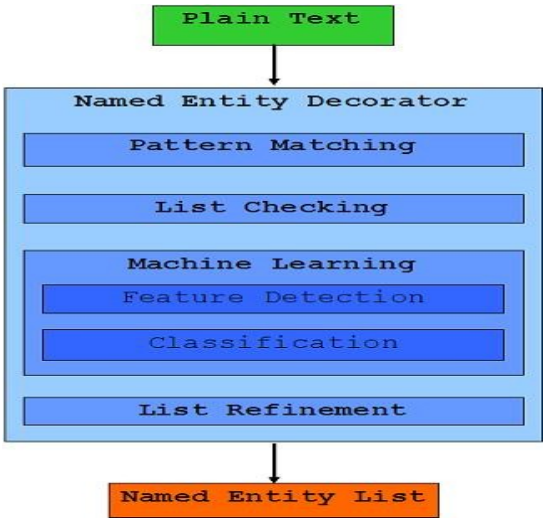


Figure 1. Architecture of NER System

### B. Background

[11] [6] The term “Named Entity” now widely used in Natural language Processing, was coined for the Sixth Message Understanding Conference (MUC-6). NER emerged as one of the subtasks of the DARPA-sponsored MUCs. At that time, MUC was focusing on Information Extraction tasks where structured information of company activities and defense related activities is extracted from unstructured text, such as newspaper articles. In defining the task, people noticed that it is essential to recognize information units like names, including person, organization and location names, and numeric expressions including time, date, money and percent expressions. Identifying references to these entities in text was recognized as one of the important sub-tasks of IE and was called “Named Entity Recognition and Classification”.
The early systems used handcrafted rule- crafted algorithms, modern systems most often resort to machine learning techniques.

[12] The computational research aiming at automatically identifying named entities in texts forma a vast and heterogeneous pool of strategies, methods and representations. One of the first research papers in the field was presented at the Seventh IEEE Conference on artificial intelligence

applications. It describes a system to extract and recognize [company] names. It relies on heuristics and handcrafted rules. It accelerated in 1996, with the first major event dedicated steady research and numerous scientific events.

[6] [12] [13] [14] A good proportion of work is devoted to the study of English but a possibly larger proportion addresses language independence and multilingualism problems. It has achieved a significant accuracy for European languages such as German, Spanish, Dutch, English, French. Japanese and Chinese have also been studied under the research. But the same task for Indian languages is lagging far behind due to various intricacies such as missing capitalization information, lack of formal and large list of gazetteers which makes pre-processing inefficient. The annotated corpora, name dictionaries, good morphological analyzers, POS taggers etc. are not yet available in the required measure. Indian languages also have the problem of disambiguation of common nouns from proper nouns. A fairly large number of frequently used words can also be used as named entities. For example, words like Vivek, Kamal, Deepak etc. which are proper nouns can also confer some meaning and can be used as common nouns. Hence the disambiguation in usage of a word as common noun or proper noun using its contextual features in Indian languages is more difficult and important than that for European languages.

### *C. Approaches*

[6] [13] [15] [16] [17] The named entity recognition systems can roughly be divided into rule based systems, which use linguistic grammar based techniques, and stochastic machine learning systems. It can also be roughly divided into three categories namely hand- crafted or automatically acquired rules or finite state patterns, look up from large name lists or other specialized resources, data driven approaches exploiting the statistical properties of the language (statistical models).

The earliest work in NER involved hand crafted rules based on pattern matching. In the second approach, the NER system recognizes only the named entities stored in its lists, also called gazetteers. This approach is simple, fast, language independent and easy to re-create the lists. However, named entities are numerous and are constantly evolving. Ststistical models have proved to be quite effective. These treat named entity recognition as a sequence tagging problem, where each word is tagged with its entity type if it is part of an entity.

The broad division of approaches to NER falls into two categories:

1. The Handcrafted Approach

i. List lookup approach: NER system uses gazetteer to classify words. Suitable lists are to be created. It is simple, fast and language independent. It is also easy to retarget as only lists are to be created. But it has disadvantage of having to maintain the gazetteer list. It cannot resolve ambiguity.

ii. Linguistic approach: NER system uses some language based rules manually written by linguists and other heuristic to classify words. It needs rich and expressive rules and gives good results. The main disadvantage is that these require huge experience and grammatical knowledge of the particular language or domain, not easily portable and has high acquisition cost. It is very specific to the target data.

2. Machine Learning based approach / Automated Approach

i. Supervised approach: It involves using a program that can learn to classify a given set of labeled examples that are made up of the same number of features. Each example is thus represented with respect to the different feature spaces. This approach requires preparing labeled training data to construct a statistical model, but it cannot achieve a good performance without a large amount of training data, because of data sparseness problem.

ii. Unsupervised approach: In this approach an unsupervised model learns without any feedback. In this learning, the goal of the program is to build representations from data. These representations can then be used for data compression, classifying, decision making, and other purposes. It is not a very popular approach for NER and the systems that do use unsupervised learning are usually not completely unsupervised.

Various machine learning approaches are:

1. Hidden Markov Models: It is a generative model. The model assigns a joint probability to paired observation and label sequence. Then the parameters are trained to maximize the joint likelihood of training sets.

$$P(X,Y) = \prod_i P(X_i, Y_i) P(Y_i, Y_{i-1})$$

It uses forward- backward algorithm, Viterbi algorithm and estimation- Modification method for modeling. Its advantage is that its basic theory is elegant and easy to understand. Hence it is easier to implement and anayze. It uses only positive data, so they can be easily scaled. Its disadvantage is that in order to define joint probability over observation and label sequence HMM needs to enumerate all possible observation sequence. Hence it makes various assumptions about data like Markovian assumption i.e. current label depends only on the previous

label. Also it is not practical to represent multiple overlapping features and long term dependencies. Number of parameters to be evaluated is huge. So it needs a large set for training.

2. Maximum Entropy Markov Models: It is a conditional probabilistic sequence model. It can represent multiple features of a word and can also handle long term dependency. It is based on the principle of maximum entropy which states that the least biased model which considers all known facts is the one which maximizes entropy. Each source state has an exponential model that takes the observation feature as input and output a distribution over possible next state. Output labels are associated with states. The advantage is that it solves the problem of multiple feature representation and long term dependency issue faced by HMM. It has generally increased recall and greater precision. The main disadvantage is that it has Label Bias Problem. The probability transition leaving any given state must sum to one. So it is biased towards states with lower outgoing transitions.

3. Conditional Random Field:
It is a type of discriminative probabilistic model. It has all the advantage of MEMMs without the label bias problem. CRFs are undirected graphical models (also know as random field) which is used to calculate the conditional probability of values on assigned output nodes given the values assigned to other assigned input nodes. Random field: Let G = (Y, E) be a graph where each vertex YV is a random variable. Suppose P(Yv | all other Y) = P(Yv | neighbors(Yv)), then Y is a random field.
Let X = random variable over data sequences to be labeled Y = random variable over corresponding label sequence. "**Definition** Let G = (V,E) be a graph such that

**Y** = (**Y**v) v∈V , so that **Y** is indexed by the vertices of G.

Then (**X**,**Y**) is a conditional random field in case, when conditioned on **X**, the random variables **Y**v obey the Markov Property with respect to the graph: P(**Y**v |**X**,**Y**w,

w ≠ v) = P(**Y**v |**X**,**Y**w, w ~ v), where w ~ v means that w

and v are neighbors in G."

$$\exp\Big(\sum_j \lambda_j t_j(y_i - 1, y_i, x, i) + \sum_k \mu_k s_k(y_i . x, i)\Big)$$

## III. NER FOR PUNJABI LANGUAGE

Punjabi is the language of Punjab, spoken mainly in Northern parts of India. Punjabi is highly inflectional and agglutinating language providing one of the richest and most challenging sets of linguistic and statistical features resulting in long and complex word forms. Each word in Punjabi is inflected for a large number of word forms. It is primarily a suffixing language. An inflected word starts with a root and may have several suffixes added to the right. It is a free word order language.

Punjabi, like other Indian languages, is a resource poor language- annotated corpora, name dictionaries, good morphological analyzers, POS taggers are not yet available in the required measure. Although Indian languages have a very old and rich literary history, technological developments are of recent origin. Web sources for name lists are available in English, but such lists are not available much in Punjabi.

### *A. Approach*

The approach being used for NER for Punjabi language in our experiment is 'Hybrid approach'. The hybrid approach is combination of the 'rule based' and the 'list look up' approaches. Number of language dependent rules are formed to implement rule based approach. And various gazetteer lists are prepared for the list look up approach.

As large corpus for Punjabi is not yet available, so it is not possible to train a system to generate rules automatically. That is why, these approaches are being used for the experiment.

### *B. Application*

The NER for Punjabi language being done in this experiment is used for the TOPIC TRACKING IN PUNJABI LANGUAGE in our another experiment. Topic tacking task is to detect news of a known topic, by monitoring a stream of news stories and finding out those which discuss the same topic described by a few positive samples. That is, the system determines whether two Punjabi news documents describe the same topic or not. The various NER features extracted in this experiment are used in topic tracking by expressing them in the form of collection of event vectors. The event vectors representing the two news documents are compared to match by at least a predefined threshold value in order to track the same topic or event.

### *C. Features Extracted*

The name entities being extracted for Punjabi language in our experiment include Name, Time/ Date, Location, Designation, Organization. Name includes the prefix, first name, middle name, last name and then forming the complete name of a person. Time/date include date, month, week day, and year. Location refers to any location name within the document. Designation includes various roles or designation names. Organization refers to name of an organization. The

NE's extracted are language dependent features and language independent features.

i. Language Independent features

The language independent features extracted are:

a. Context word feature: preceding and following words of a particular word.

b. Presence of digits: may represent date, time, month, and year.

c. Complete word: if a word is a complete word or it is a part of another word.

ii. Language Dependent features

A number of rules specific for Punjabi language are formed to extract the language dependent features.

Date/time rules:

- Any format of the form dd/mm/yyyy, dd-mm-yyyy or dd.mm.yyyy, is extracted as date. E.g. 30/11/2010, 30-11-2010 or 30.11.2010
- Any format of the form yyyy-yy or yyyy-yyyy is extracted as year. E.g 1998-99
- When month name is found, it is extracted as month.
  - The previous word is checked, if it is of the form dd(<=31), then extracted as date. E.g if it is 31 ਜਨਵਰੀ, then 31 is extracted as date entity.
  - The next word is checked, if it is of the form yyyy, then extracted as year. e.g if it is ਜਨਵਰੀ 2011, then '2011' is extracted as year entity.
- When week day is found, it is extracted.
- If Punjabi word '**ਸੰਨ**' (sann) or '**ਸਾਲ**'(sāl) is found followed by three ([1-9][0-9][0-9]) or four([1-2] [0-9][0-9][0-9]) digits, then it represents year.
- If Punjabi word **ਈ**. (ī.) or **ਵਿਚ** is found, its previous word checked for three ([1-9][0-9][0-9]) or four([1-2][0-9][0-9][0-9]) digits, then it represents year.

b. Designation rule:

- When designation name found, it is extracted.
- For a two word designation such as '**ਡਿਪਟੀ**'(ḍipṭī) as first word, next word is checked for designation. If it is, then the word and the next word are collectively taken as designation. e.g. **ਡਿਪਟੀ ਕਮਿਸ਼ਨਰ** (ḍipṭī kamishanar)

c. Organization rule:

- When an organization suffix such as **ਕੰਪਨੀ** (kampnī), **ਕਮੇਟੀ** (kamēṭī ), **ਕਲੱਬ** (kalabb), **ਦਲ**(dal), **ਬੋਰਡ** (bōraḍ), **ਵਿਭਾਗ**(vibhāg), **ਆਰਗੇਨਾਈਜੇਸ਼ਨ** (ārgēnāījēshan), **ਐਸੋਸੀਏਸ਼ਨ** (aisōsīēshan), **ਯੂਨੀਅਨ**(yūnīan) etc is found, then its previous word and the suffix are collectively extracted as organization.

d. Name rules:

- Prefix rule:
  - When some prefix is found, its next word is taken as first name.
  - Word next to first name is checked for middle name or last name. If it is, then both are concatenated.
  - Word next to middle name is checked for last name. If it is, again complete name is formed by concatenation.
- Middle name rule:
  - When Middle Name is found the previous word is taken as First Name.
  - Also the word after the Middle Name is checked whether it is Last Name or not. If it is Last Name, the First Name, Middle Name and Last Name are concatenated and complete name is formed. Otherwise just First Name and Middle Name is concatenated as name.
- Last name rule:
  - When Last Name is found the previous word is checked if it is Middle Name or not. If it is not Middle Name then it is taken as First Name.
  - If it is Middle Name then the word before the Middle Name is taken as First Name. The Founded First Name, Middle Name and Last Name are concatenated to form name.

Location rule:

- ਵਿਖੇ rule: when Punjabi word ਵਿਖੇ is found, its previous word as extracted as a location name.
- Location suffixes 'ਪੁਰ' or 'ਗੜ੍ਹ' found, then complete word is extracted as location name.

*D. Gazetteer Lists*

- Names of months
- Days of a week
- Prefix for name
- Designation names
- Organization suffix
- Middle names
- Last names
- Stop words
- Location names

*E. Methodology*

First, all stop words are removed from the document, because they slow down the process as they are large in number and are of no use. Then rules are implemented which use the gazetteers lists to extract Names, time/ date, location, designation, organization from the document.

*F. Evaluation Metrics*

The performance of NER system is measured using precision(P), recall(R) and F-measure.

The precision measures the number of correct NEs, obtained by NER system, over the total number of NEs extracted by NER system.
P= number of correct NEs/total number of NEs

The recall measures the number of correct NEs, obtained by NER system over the total number of NEs in a text that have been used for testing.

R= number of correct NEs/total number of NEs in a text

The F-measure represents harmonic mean of precision and recall.

F= 2RP/R+P

*G. Implementation Details*

The system has been implemented using vb.net platform and gazetteers lists are stored as tables in the database. The experiment required input documents which contain name entities such as person names, location names, designations, organization names, time/ date. Such test documents taken from following Punjabi web sources such as likhari.org, punjabispectrum.com , europevichpunjabi.com, quamiekta.com, ajitweekly.com, sahitkar.com onlineindian.net, europesamachar.com, jagbani.com, parvasi.com.

*H. Experimental Results*

The experimental results reported in table show that

| NE Class | P(%) | R(%) | F(%) |
|---|---|---|---|
| Person | 74.52 | 62.86 | 65.67 |
| Location | 91.52 | 92.89 | 91.25 |
| Organization | 90.27 | 90.10 | 88.77 |
| Designation | 98.84 | 87.09 | 91.98 |
| Date/Time | 94.79 | 89.79 | 91.75 |
| Total | 89.98 | 84.55 | 85.88 |

The system shows good results for location, organization, designation, date/ time NER's. The results for person name are not good as compared to other NEs.

## IV. CONCLUSION & FUTURE SCOPE

Not much work has been done in NER in Punjabi and other Indian languages. In this paper, we have reported our work on Name Entity Recognition for Punjabi. We have prepared a 'hybrid system', with the combination of two NER approaches, i.e. 'rule based approach' and 'list look up approach'. A number of language dependent rules are formed to extract language dependent features for Punjabi and number of language independent rules are formed which can be used for any other language also. The list look up approach uses the gazetteer lists created to extract various NEs. Hence, NEs such as date/ time, location, person name, organization, designation for Punjabi language are extracted. The system shows good results for date/time, location, organization, designation and low results for person name as compared to other NEs.

Future works include forming new rules to improve the existing results. As many first names in Punjabi are also common nouns, this limitation lowers the performance of the system. This issue can be considered to improve the system. More name entities, apart from these, can be extracted such as title, monetary expressions, measurement expressions etc.

## REFERENCES

[1] http://en.wikipedia.org/wiki/Natural_language_processing
[2] Gobinda G. Chowdhury ,Dept. of Computer and Information Sciences ,University of Strathclyde, Glasgow G1 1XH, UK

[3] www.cnlp.org/publications/03nlp.lis.encyclopedia

[4] Sujeet Kumar, (2008), " Named Entity Recognition for Hindi", Indian Institute of Technology, Kanpur

[5] Tzonhan Tsai, Shihung Wu, Chengwei Lee, Chengwei Shih, and Wenlian Hsu, "Mencius: A Chinese Named Entity Recognizer using the Maximum Entropy based Hybrid Model", International Journal of Computational Linguistics of Chinese Language Processing, Vol. 9; Nov. 1, 2004

[6] Named Entity Recognition for Telugu

[7] Andrew McCallum and Wei Li, "Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons:, in 7th Conference on Natural Language Learning( CoNLL)

[8] Wei Li and Andrew McCallum, "Rapid Development of Hindi Named Entity Recognition Using Conditional Random Fields and Feature Induction", in ACM Transactions on Asian language information Processing, 2003

[9] ZhenzhenKou, William W. Cohen,(2005) "High-Recall Protein Entity Recognition Using a Dictionary", in 13th Annual International Conference on Intelligent Systems for Molecular Biology

[10] Mohammad Hasanuzzaman, Asif Ekbal and Sivaji Bandyopadhay, (2009), " Maximum Entropy Approach for Named Entity Recognition in Bengali and Hindi", Academy Publisher, International Journal of Recent Trends in Engineering, Vol. 1, No. 1.

[11] R. Grishman, Sundheim,(1996), " Message Understanding Conference-6:A Brief History", Proceedings of International Conference on Computational Linguistics.

[12] Lisa F. Rau, (1991), "Extracting Company Names from Text", IEEE, Proceedings of Conference on Artficial Intelligence Applications of IEEE.

[13] Awaghad Ashish Krishnarao, Himanshu Gahlot, Amit Srinet, D.S. Kushwaha, (2008), " A Comparative Study of Named Entity Recognition for Hindi Using Sequential Learning Algorithms", IEEE, International Advance Computing Conference(IACC 2009)

[14] Sujan Kumar Saha, Sudeshna Sarkar, Pabitra Mitra, (2008), "Gazetteer Preparation for Named Entity Recognition in Indian Languages", the 6th workshop on Asian Language Resources

[15] Alireza Mansouri, Lilly Suriani Affendey, Ali Mamat, (2008), " A New Fuzzy Support Vector Machine Method for Named Entity Recognition", IEEE, International Conference on Computer Science and Information Technology

[16] Branimir T. Todorovic, Svetozar R. Rancic, Ivica M. Markovic, Edin H. Mulalic, Velimir M Ilic,(2008), " Named Entity Recognition and Classification using Context Hidden Markov Model",IEEE,9th Symposium on Neural Network Applications in Electrical Engineering

[17] A Hybrid Approach for Named Entity Recognition in Indian Languages

[18] Xu-Dong Lin, Hong Peng, Bo Liu, (2006), "Chinese Named Entity Recognition Using Support Vector Machines", IEEE, Proceedings of the 5th International Conference on Machine Learning and Cybernetics, Dalian, 13-16 August 2006G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

AUTHORS PROFILE

**Kamaldeep Kaur** has done M.E. in Computer Science & Engineering at University Institute of Engineering & Technology, Panjab University Chandigarh. She has done B.Tech. in Computer Science & Engineering from Guru Nanak Dev Engineering College, Ludhiana in 2008.She is among university toppers. She secured 82% Marks in B.Tech. Kamaldeep is devoting her research work in field of Natural Language processing. She is also a merit holder in 10th and 12th classes of Punjab School education board

**Vishal Gupta** is Lecturer in Computer Science & Engineering Department at University Institute of Engineering & Technology, Panjab University Chandigarh. He has done M.Tech. in computer science & engineering from Punjabi University Patiala in 2005. He secured 82% Marks in M.Tech. He did his B.Tech. in CSE from Govt. Engineering College Ferozepur in 2003. He is also pursuing his PhD in Computer Sc & Engg. Vishal is devoting his research work in field of Natural Language processing. He has developed a number of research projects in field of NLP including synonyms detection, automatic question answering and text summarization etc. One of his research papers on Punjabi language text processing was awarded as best research paper by Dr. V. Raja Raman at an International Conference at Panipat. He is also a merit holder in 10th and 12th classes of Punjab School education board. in professional societies.